

Backtesting VaR models: Quantitative and Qualitative Tests

Carlos Blanco and Maksim Oks

This is the first article in a two-part series analyzing the accuracy of risk measurement models. In this first article, we will present an overview of backtesting methods and point out the importance of conducting regular backtests on the risk models being used. In the second article, we will present an alternative to measuring VaR using a top-down or “macro” approach as a complementary tool to traditional risk methodologies.

Should risk models be accurate?

Firms that use VaR as a risk disclosure or risk management tool are facing growing pressure from internal and external parties such as senior management, regulators, auditors, investors, creditors, and credit rating agencies to provide estimates of the accuracy of the risk models being used.

Users of VaR realized early that they must carry out a cost-benefit analysis with respect to the VaR implementation. A wide range of simplifying assumptions is usually used in VaR models (distributions of returns, historical data window defining the range of possible outcomes, etc.), and as the number of assumptions grows, the accuracy of the VaR estimates tends to decrease.

As the use of VaR extends from pure risk measurement to risk control in areas such as VaR-based Stress Testing and capital allocation, it is essential that the risk numbers provide accurate information, and that someone in the organization is accountable for producing the best possible risk estimates. In order to ensure the accuracy of the forecasted risk numbers, risk managers should regularly backtest the risk models being used, and evaluate alternative models if the results are not entirely satisfactory.

VaR models provide a framework to measure risk, and if a particular model does not perform its intended task properly, it should be refined or replaced, and the risk measurement process should continue. The traditional excuse given by many risk managers that “VaR models only measure risk in normal market conditions” or “VaR models make too many wrong assumptions about market or portfolio behavior” or “VaR models are useless” should no longer be taken seriously, and risk managers should be accountable to implement the best possible framework to measure risk, even if it involves introducing subjective judgment into the risk calculations. It is always better to be approximately right than exactly wrong.

Determining the accuracy of VaR models

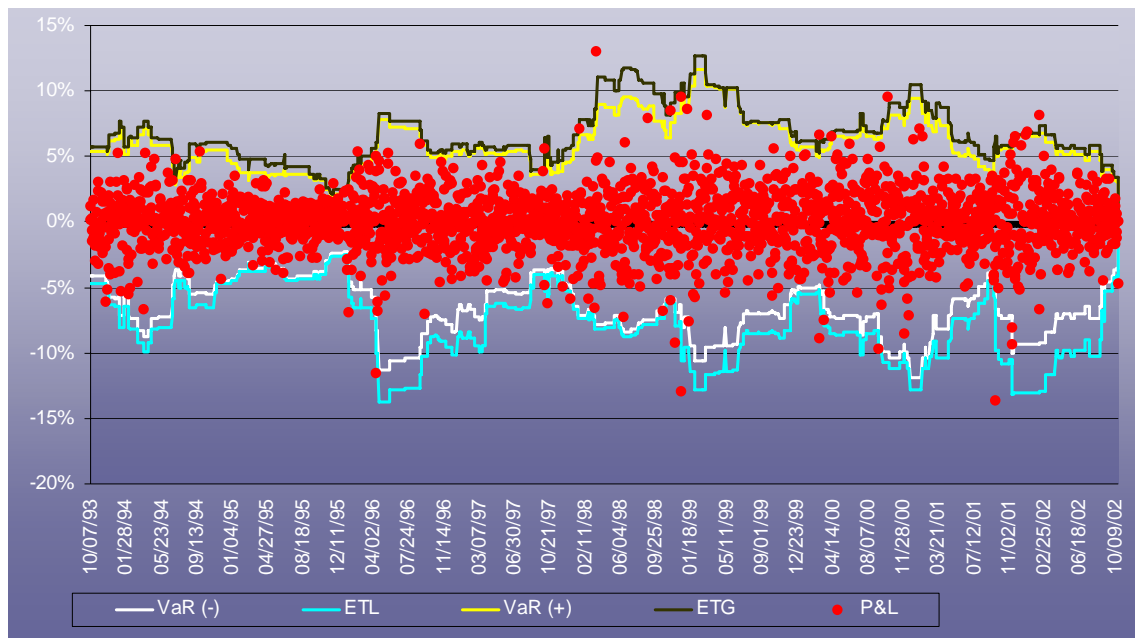
How can we assess the accuracy and performance of a VaR model? To answer this question, we first need to define what we mean by “accuracy.” By accuracy, we could mean:

- How well does the model measure a particular percentile of or the entire profit-and-loss distribution?
- How well does the model predict the size and frequency of losses?

Many standard backtests of VaR models compare the actual portfolio losses for a given horizon vs. the estimated VaR numbers. In its simplest form, the backtesting procedure consists of calculating the number or percentage of times that the actual portfolio returns fall outside the VaR estimate, and comparing that number to the confidence level used. For example, if the confidence level were 95%, we would expect portfolio returns to exceed the VaR numbers on about 5% of the days.

Backtesting can be as much an art as a science. It is important to incorporate rigorous statistical tests with other visual and qualitative ones.

Simple Backtesting: VaR estimates vs. P&L



The simplest backtest consist of counting the number of exceptions (losses larger than estimated VaR) for a given period and comparing to the expected number for the chosen confidence interval.

A more rigorous way to perform the backtesting analysis is to determine the accuracy of the model predicting both the frequency and the size of expected losses. Backtesting Expected Tail Loss (ETL) or Expected Tail Gain (ETG) numbers can provide an indication of how well the model captures the size of the expected loss (gain) beyond VaR, and therefore can enhance the quality of the backtesting procedure.

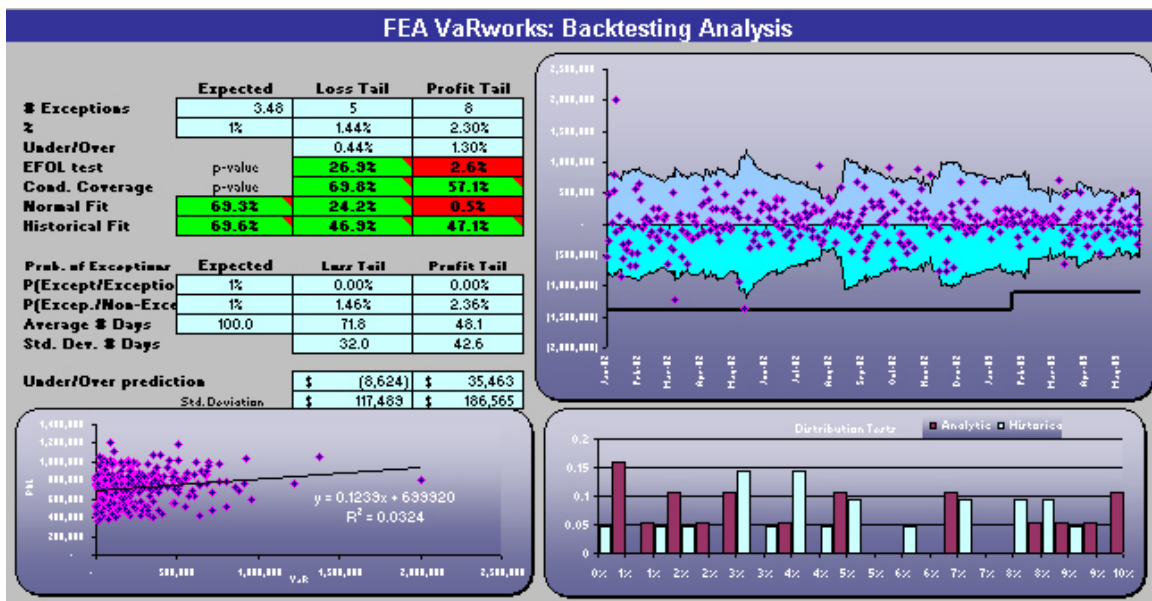
* If we do not have ETL, VaR(+) and ETG data, we can perform the analysis with VaR data exclusively, but we would have limited information to extract conclusions.

Quantitative tests:

Statistical tests help us check whether the risk model is accurately capturing the frequency, independence or magnitude of exceptions, which are defined as losses (gains) exceeding the VaR estimate for that period.

When we test a certain hypothesis in statistics, we can make two types of errors: Type I errors occur when we reject the model which is correct, while type II errors occur when we fail to reject (that is incorrectly accept) the wrong model. It is clear that in risk management, it can be much more costly to incur in type II errors, and therefore we should impose a high threshold in order to accept the validity of any risk model.

The implications for the choice of the confidence level for the VaR calculations, is that the larger the confidence level for the VaR estimates, the fewer the number of “exceptions” and therefore, it will be more difficult to validate the model. If we choose a 95% level, that means that we will be able to observe more “exception” points than the 99% level, and therefore we will have a better test of the model accuracy.



Many statistical tests are based on the frequency and time dynamics of exceptions. We briefly discuss the most common ones:

Test of Frequency of Tail Losses or Kupiec test.

Kupiec’s (1995) test attempts to determine whether the observed frequency of exceptions is consistent with the frequency of expected exceptions according to the VaR model and chosen confidence interval. Under the null hypothesis that the model is “correct”, the

number of exceptions follows a binomial distribution. The probability ¹ of experiencing x or more exceptions if the model is correct is given by:

$$\Pr(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where x is the number of exceptions, p is the probability of an exception for a given confidence level, and n is the number of trials.

If the estimated probability is above the desired “null” significance level (usually 5% - 10%), we accept the model. If the estimated probability is below the significance level, we reject the model and conclude that it is not correct. We can conduct this test for loss and gain exceptions to determine how well the model predicts the frequency of losses and gains beyond VaR numbers.

Conditional Coverage of Frequency and Independence of Tail Losses (Christoffersen test).

The Kupiec tests only focuses on the frequency of exceptions, and ignores the time dynamics of those exceptions. VaR models assume that exceptions should be independently distributed over time. If the exceptions exhibited some type of “clustering”, then the VaR model may fail to capture P&L variability under certain conditions, which could represent a potential problem down the road.

The main contribution of this approach is its ability to test sub-hypothesis regarding the frequency and independence of exceptions, and the joint hypothesis that the VaR model has the right frequency of independent exceptions.

An added benefit of conducting these type of tests is that it generates some additional useful information such as the conditional probabilities of experiencing an exception followed by an exception in the risk model, and the average number of days between exceptions.

Problems with statistical tests and possible solutions

The standard tests that focus on frequency and independence of exceptions have the problem that they are weak and often fail to properly exclude the null hypothesis and therefore incur in a Type II error. Besides the low power of the standard tests, there is another issue: The “true” null probability is not known. Therefore when we use a test we do not know if we may be accepting a wrong model or rejecting a good one because we may be using the wrong null probability.

¹ This probability, also known as p-value, is the probability of getting a sample, which is even less likely than the sample we actually have, given that the null hypothesis is true.

However, there is a possible solution suggested by Dowd (2003). We can use a bootstrapping mechanism to construct a sample of null hypothesis probabilities that can then be used as a back-testing input. Bootstrapping involves creating alternative samples by drawing observations from our original sample of VaR and P&Ls, and replacing the observation in the sample pool after it has been drawn. We can repeat this process as many times as we wish, and create alternative samples from which we can estimate the p-values for the Kupiec and Christoffersen tests. That way, we have a “sample of sample estimates” from which we can construct confidence intervals for those parameters.

The bootstrapped values can provide a confidence band around the results of the statistical tests, and as Dowd (2003) points out, “it is better to conclude that we are not confident enough about the model one way or another, and be right, than to come to the confident conclusion that we are wrong. The bootstrap back-test procedure allows us to take a leaf from the Scottish justice system: sometimes the correct verdict given the evidence is not “guilty” nor “not guilty”, but “not proven”.

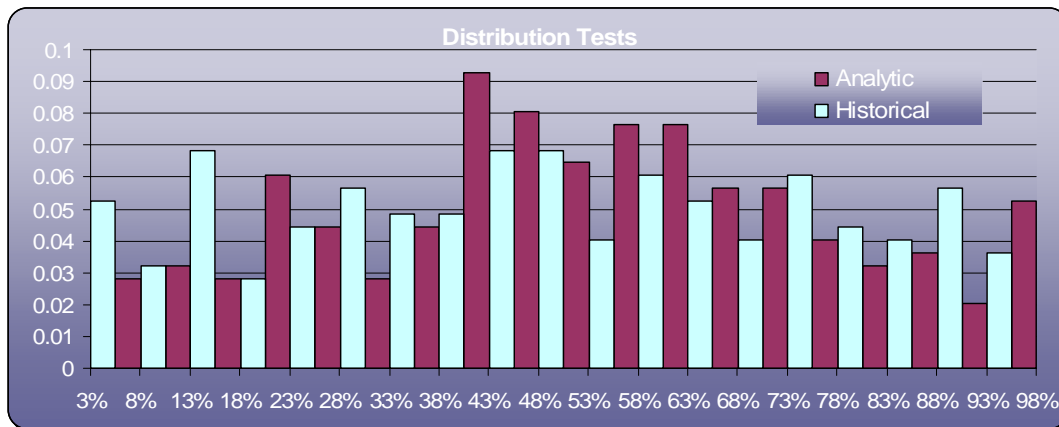
Tests based on the entire distribution or the tail of the distribution

In addition to backtesting the traditional interval and point risk measures such as VaR and ETL, one could also be interested in backtesting how well the model predicts the entire distribution of profits and losses. This has an added benefit of further rejecting bad models.

In this approach, forecasts at many quantiles are compared to the realized data and the probability of observing a return below the actual is calculated. We denote this so-called transform probability by p_{t+1} :

$$p_{t+1} \equiv F_t(\text{Return}_{t+1})$$

If the risk model is correct, then the time series of observed probabilities p_{t+1} should be independent and identically distributed (i.i.d.) as a uniform (0,1) variable. One can then perform a graphical analysis by simply constructing a histogram of these probabilities and checking that it looks reasonably flat.



In risk management, it is usually important that the model forecasts the tail of the distribution correctly and not its interior, which characterizes small return disturbances. However, the approach described above can easily be generalized to testing only the tails of the distribution see Berkowitz (2000).

This information about the distribution of p_{t+1} can be converted from qualitative, graphical analysis into statistically testable hypotheses. Berkowitz (2000) describes two powerful Likelihood Ratio (LR) tests for testing the full distribution and the tail of the distribution that are based on this approach.

A more general framework for Backtesting: Forecast Evaluation Approaches

A more general framework involves specifying a set of rules that will define the “accuracy” of the model according to a risk manager, and design method of evaluating the model’s performance according to those rules.

These type of approaches are not formal hypothesis tests, but instead involve specifying a loss function that reflects the preferences of the institutions of some models over others. This “loss function” approach can be a useful supplement to these more formal statistical methods and provides a way to define the institution’s criteria to define an “accurate” model. For example, we can design a loss function in which the modeler can weight the penalties to assign to exceptions given their frequency, magnitude or time dependencies and compare them with expected tail loss numbers. The main benefit of this type of analysis is that it provides a measure of relative performance that can be used for "backtesting" different models. For more information on these approaches, see Dowd (2002).

A very useful metric is to compare the exception sizes vs. the Expected Tail Loss and Expected Tail Gain to determine whether the model over or under-predicts the size of expected losses and gains.

Conclusion

Even though it seems like the first step before using a risk model to make decisions is to check its accuracy, relatively little attention has been given to this topic by practitioners or academics. In this article, we have presented a set of qualitative and quantitative tools to conduct this analysis on a regular basis. In the next article, we will discuss an alternative way to calculate VaR and ETL by using a “top-down,” or “macro” approach that can easily be backtested “on the fly,” without the need to wait for months or years after a risk system is implemented.

Bibliography

Berkowitz J. (2000) Testing Density Forecasts, with Applications to Risk Management. Graduate School of Management, University of California, Irvine.

Christoffersen, P. (1998) Evaluating Interval Forecasts, *International Economic Review*, 39, 841-862

Dowd K. (2002). A Bootstrap back-test. October. *Risk*. 93-94

Dowd, K. *Measuring Market Risk*. John Wiley and Sons

Kupiec, P., Techniques for verifying the accuracy of risk management models. *Journal of Derivatives* 3. pages 73-84

Carlos Blanco, Ph.D. is VP, Risk Solutions at FEA, a Barra Company.

Maksim Oks, Ph.D. is Financial Engineer and has been in charge of research and development efforts of the Backtesting Module that will soon be released by FEA.